第62卷 第9期 2018年5月

基于SKOS的学术期刊文本资源多粒度语义标注方法研究*

夏立新 郑路 张玉晨 翟姗姗 孙晶琼
华中师范大学信息管理学院 武汉 430079

摘要：[**目的/意义**]针对学术期刊文本资源语义标注仍存在的通用本体难以构建、标注粒度单一两大问题,提出基于SKOS的学术期刊多粒度语义标注方法,从而进一步推进语义标注的应用发展,更好满足用户的多粒度学术信息需求。[**方法/过程**]在对《中国汉语主题词表》进行SKOS描述的基础上,以学术期刊文本资源为对象,实现其多粒度语义标注,并通过实证研究验证该方法的可行性。[**结果/结论**]利用SKOS实现对学术期刊文本资源进行多粒度语义标注,较之当前学术检索系统中的标注结果,在"查全""查准""内部特征检索入口""检索结果反馈形式"4个方面具有一定优势。

关键词：语义标注 多粒度 SKOS 叙词表 学术期刊
分类号：G250
DOI:10.13266/j.issn.0252-3116.2018.09.015

1 引言

学术期刊是学术成果展示与交流的重要平台,随着数字化的不断发展,海量的的学术信息在满足用户研究需求的同时,也带来了严重的信息过载压力,使用户难以高效地提取出所需的学术信息[1]。同时,不同来源的学术期刊标准和规范不同,给学术期刊的传播和共享带来了不便。语义标注的出现为学术期刊提供了新的知识组织方法,通过对原始数据作标记,使其具有语义信息,不仅人可以理解,而且使机器也可以理解[2],大大提高了学术期刊的检索与利用效率。

目前基于本体的语义标注方法使用最为广泛,但直接构建一个领域本体往往消耗较大,更需要领域专家的协助。不同本体之间也常常存在异构问题,难以通用和融合。因此,如何构建一个真正通用的本体成为了基于本体的语义标注方法的一大瓶颈。万维网联盟(W3C)于2004年公布的简单知识组织系统(simple knowledge organization system, SKOS)标准提供了对受控词表的知识管理及语义处理方案[3],使用SKOS化的受控词表进行语义标注成为了解决这一问题的一种可能方案。此外,目前针对学术期刊的标注更多地使用粗粒度的标注方法,学术期刊内部所蕴含的大量信息无法被进一步检索、过滤和提取;为解决这一问题,也有学者进行了学术期刊的细粒度标注,将标注单位深入到最小知识单元。无论粗粒度或细粒度的标注方法都仅能实现单一粒度的知识组织,但用户的学术信息需求却常常呈现多粒度性。因此,单一粒度的标注将无法满足用户的学术信息需求,多粒度的标注则成为了满足用户多粒度的学术信息需求的基础。

为了尝试进一步推动语义标注的应用发展,更好地满足用户的多粒度学术信息需求,本文以SKOS资源描述框架与相关技术为基础,探讨学术期刊文本资源的多粒度语义标注的实现途径。

2 相关研究

2.1 SKOS研究现状

SKOS是以RDF资源描述框架为基础,用来描述受控词表的基本结构和概念的标准语言。受控词表可通过SKOS资源描述框架转换为与RDF、OWL兼容的概念模型,实现语义化的信息资源共享。

自SKOS资源描述框架发布以来,国内外对SKOS

* 本文系国家社会科学基金重大项目"基于多维度聚合的网络资源知识发现研究"(项目编号:13&ZD183)和国家社会科学基金青年项目"面向语义出版的数字图书馆资源多维度聚合研究"(项目编号:15CTQ007)研究成果之一。
作者简介:夏立新(ORCID:0000-0002-4162-2282),教授,博士生导师;郑路(ORCID:0000-0001-5870-9803),博士研究生;张玉晨(ORCID:0000-0003-1451-7871),硕士研究生,通讯作者,E-mail:brettzhang_edu@163.com;翟姗姗(ORCID:0000-0002-2787-0183),副教授;孙晶琼(ORCID:0000-0002-7074-9602),硕士研究生。
收稿日期:2017-10-22 修回日期:2018-01-17 本文起止页码:123-133 本文责任编辑:易飞

的研究主要集中于受控词表的SKOS转化问题。目前英文受控词表的SKOS化已有不少成功的例子,在W3C的SKOS官方网站中Datasets页面已共享了多达39个SKOS化的受控词表[4],但遗憾的是其中并无中文受控词表。对于中文受控词表来说,学者们的研究主要集中在《汉语主题词表》或《中国分类主题词表》的SKOS转换上。范炜提出了利用SKOS构造机器可理解的知识组织体系,并以叙词表为例进行了实例研究[5]。贾君枝针对《汉语主题词表》对SKOS的内容及结构作明确描述,完成了《汉语主题词表》的SKOS描述示范[6]。此外,张士男等设计了《中国科学院图书馆图书分类法》的SKOS描述方案[7]。

目前已SKOS化的受控词表虽然已有一定成果但还未能得到广泛应用,相关研究并不集中。J. Pastor-sanchez等将SKOS方案与语义网中其他受控词表的表示方案进行了比较分析,并最终认为SKOS是叙词表的最佳语义描述方案[8]。王茜等从宏观上讨论了使用SKOS对语义网进行知识组织的模型,通过对SKOS模型中类与属性的扩展增强了对知识的描述能力[9]。熊太纯详细分析了SKOS对网络环境下信息资源进行标引的可行性[10]。

综上所述,目前SKOS资源描述框架的研究主要集中于受控词表的语义描述方面,并且受控词表SKOS的描述成果已相当丰富,但其后续的应用研究相对欠缺,更缺少实践成果。

2.2 学术期刊语义标注研究现状

朱嘉贤等将语义标注总结为"利用本体技术将信息资源中的概念、属性、关系等语义信息标注为计算机可理解的元数据,实现标注信息与资源的关联。"[11]

目前,针对于学术期刊文本资源所采用的语义标注方法主要是基于本体技术实现的。在理论层面,魏墨济等提出一种基于领域本体的学科专业文档的语义标注方法[12];冷伏海等综合运用语义标注技术、规则抽取技术以及正则表达式技术,提出一种抽取学术文献中涉及的具体理论等学术信息的方法[13];英国谢菲尔德大学研发的文本工程通用框架GATE是基于多本体的语义标注方法方面的突出代表,但不足的是该平台的多本体之间没有建立映射关联,难以互联互通[14]。此外也有很多学者提出了很多更为具体的基于本体的学术期刊文本资源语义标注优化方法,但大部分方法还处于尝试阶段。在实践层面,DBpedia项目是文本资源语义标注的典型代表,该项目从Wikipedia(维基百科)的词条里抽取出结构化的信息,并将其他资料集关联到维基百科,最终将数据集以关联数据(linked data)的形式发布[15]。此后,众多学者以该项目为基础,开展了学术文本资源语义标注的实践尝试,如F. Norberto等学者提出了一种基于DBpedia的协作语义标注框架,该框架充分利用了人工语义标注的优势,将基本用户操作与语义标注操作融合,同时减轻了非专家标注者的负担[16];汤怡杰等将中国科学院集成信息平台(CASIIP)与DBpedia数据集相结合,利用DBpedia内部的信息资源描述和组织形式将CASSIP中的数据信息进行语义标注,实现CASIIP平台的语义化扩展[17]。

综上,针对于学术文本资源的语义标注研究已取得了一定的成果积累,但其研究主要以学术资源出版单位为对象,多针对于整个文档或整个资源集合,从标注结构来说,并未深入到某篇文档的章节中,从标注内容来说,并未涉及到标识文档内容特征的知识单元。换言之,在对学术文本资源所进行的语义标注研究中,并未考虑更细粒度的语义标注方案[18],且缺乏一个使学术文本资源语义描述更加结构化、规范化的通用本体。这两大问题则直接影响了用户多粒度的学术信息需求,用户学术信息需求越深入、越细致,越无法有效定位自身所需的信息资源。因此,本研究在已有的研究基础上,通过对《汉语主题词表》的SKOS化,进一步实现学术期刊文本资源的多粒度语义标注。

3 SKOS在学术期刊文本资源多粒度语义标注中的应用分析

3.1 多粒度语义标注的优势

多粒度的语义标注简单来说就是将标注文档内容进行粒度划分之后,分别对每个粒度层进行语义标注,形成有层次、有结构的语义标注结果。粗粒度是对某一主题全面的描述;中粒度是对某一主题其中某一方面的描述;细粒度是对某一具体问题的描述。

多粒度标注对文档的揭示既有整体性的概括也有深入文档具体内容的描述,丰富的标注成果较之目前常使用的粗粒度或细粒度的单一粒度的标注,不论是对进一步更高层次的知识组织形式还是用户的检索反馈都能提供更大的支持。

3.2 SKOS描述叙词表的优势

目前已经有了很多基于XML和RDF的叙词表描述方案,还有一些在某些方面的替代方法,如主题图等。与其他叙词表的表示方案相比,SKOS主要具有表1所示的明显优势:

表 1　SKOS 较其他叙词表描述方案的优势

叙词表描述方案	SKOS 的相对优势
XML 词汇(ZTHES,MESH)	使用 RDF 在描述层面即可集成语义
概念地图,主题图(XTM)	基于 OWL 在逻辑层面集成语义
其他 RDF 词汇(LIMBER,CERES,ILRT)	概念描述的变化更加灵活且标准化
OWL 本体	描述和维护任务简单

3.3　《汉语主题词表》的 SKOS 描述

本文选择《汉语主题词表》的 SKOS 描述成果作为多粒度语义标注方法的标注工具。下面将对《汉语主题词表》的 SKOS 描述过程进行说明。

《汉语主题词表》中的叙词或非叙词都是作为 SKOS 概念的词汇标签进行描述的。如果 SKOS 标准语言中没有与词汇属性相对应的属性,则对 SKOS 标准语言进行定制化扩展,增加新属性,扩展部分称为 SKOSEX 语言,在对《汉语主题词表》的描述中本文需要使用部分 SKOSEX 语言进行属性描述。除概念及语义关系描述外也有一些属性是为创建词表而设立的,如增词时间、词频、词类型、编辑次数等,这些属性不需向用户展示,因此可以忽略。最终,《汉语主题词表》中的术语本文采用表 2 中的属性进行描述:

表 2　《汉语主题词表》的术语描述属性

中文叙词表中常见的词汇关系	对应的 SKOS 概念属性	说明
《中国图书馆分类法》	skos:broadMatch	《中国图书馆分类法》中所属的类号
范畴	skos:broadMatch	范畴索引中所属的范畴号
汉语拼音	skos:prefLabel xml:lang = "zh-latn"	叙词的拼音表示
英文	skos:prefLabel xml:lang = "en"	叙词的英译名称
中文	skos:prefLabel xml:lang = "zh"	叙词的中文标签
缩略语	skosex:abbreviation	叙词的缩略语,一种可选标签
同项	skos:exactMatch	两个叙词之间的等同关系
代项	skos:altLabel xml:lang = "zh"	与叙词同义的非叙词
属项	skos:broader	上位概念
分项	skos:narrower	下位概念
参项	skos:related	相关概念
族项	skos:topBroader	族首词
领词	skos:leadBroader	分词族的族首词
见代	skos:related	把叙词引入到其他相关叙词
和项	skosex:coordinationOf	组配成复合概念的一个成分概念
用和	skosex:coordinationOf	组配生成复合概念的成分概念
组代	skosex:coordinatedTo	单一叙词组配而成的复合概念
代码	skos:notation	叙词对应的某种标记符号
注释	skos:note	注释属性
历史注释	skos:historyNote	历史注释
用户评价	skosex:evaluationNote	用户评价注释

其中,xmlns:skos 表示 W3C 定义的 SKOS 标准语言,xmlns:skosex 表示对 SKOS 的扩展语言,属性 xml:lang 的语言代码由《IETF BCP47》标准定义。

如《汉语主题词表》中的叙词"情报检索"的部分 SKOS 语言描述示例见图 1。

4　基于 SKOS 的学术期刊文本资源多粒度语义标注框架设计

本文研究流程如图 2 所示。具体而言:①进行叙词表语义描述,包括语义关系的揭示;②设计学术期刊文本资源的多粒度标注过程;③结合二者实现基于 SKOS 的学术期刊文本资源多粒度语义标注;④按照基于 SKOS 的学术期刊文本资源多粒度语义标注方法实施方案,以期刊论文为例进行标注,验证该方法的可行性,并对标注结果进行评估。

基于 SKOS 的学术期刊文本资源多粒度语义标注框架可分为 3 个主要部分:

(1)叙词表向 SKOS 转化。叙词表是本方法在标注过程中选择使用的情报检索语言,在语义标注中需要先将传统的叙词表中的概念和语义关系使用 SKOS 标准描述语言表达,SKOS 化的叙词表是该方法进行语义标注的基本工具。本文选择《汉语主题词表》的

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdf="http://www.w3.org/2004/02/skos/core#">

<skos:Concept rdf:about="http://www.example.com/concepts#情报检索">

<skos:prefLabel>情报检索<skos:prefLabel>
<skos:prefLabel>information retrival<skos:prefLabel>
<skos:altLabel>文献检索<skos:altLabel>
<skos:altLabel>信息检索<skos:altLabel>
<skos:altLabel>文献信息检索<skos:altLabel>
<skos:narrower rdf:resource="http://www.example.com/concepts#检索语言"/>
<skos:related rdf:resource="http://www.example.com/concepts#检索"/>
<skos:related rdf:resource="http://www.example.com/concepts#情报检索工具"/>
<skos:related rdf:resource="http://www.example.com/concepts#查询"/>

</skos:Concept>

</rdf:RDF>
```

图 1 叙词“情报检索”的 SKOS 描述

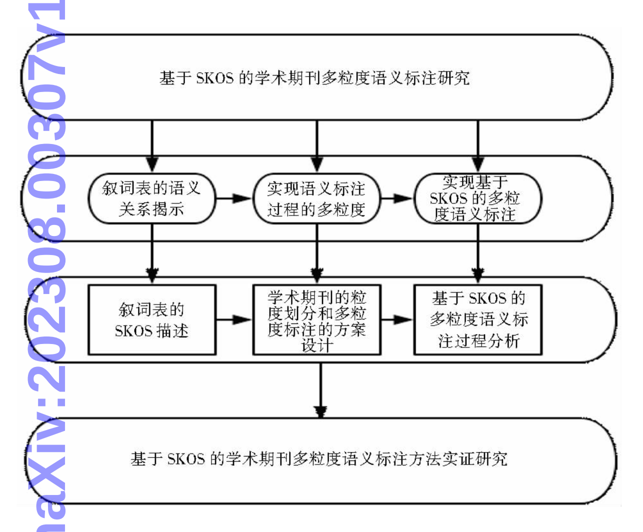


图 2 基于 SKOS 的学术期刊多粒度语义标注方法框架

SKOS 描述成果作为多粒度语义标注方法的标注工具。《汉语主题词表》的 SKOS 描述方案已于 3.3 小节说明。

(2)学术期刊文本资源的多粒度处理及标注词选取。在进行语义标注之前,标注对象需要完成 3 个基本处理。首先,对标注对象进行多粒度的层次构建,通过学术期刊文本资源的粒度划分将学术期刊内容按照不同大小的粒度单位构建形成等级树状结构。其次,对学术期刊内容依照粒度划分结果进行多粒度的分词。最后,在多粒度分词的结果上进行多粒度标注候选词的重要性计算,即构建标注词评价指标,计算各词得分后依据得分高低选取当前粒度的标注词。

(3)基于 SKOS 的学术期刊文本资源多粒度标注及结果生成。通过计算选择出合适的标注词后使用 SKOS 描述的叙词表对学术期刊文本资源进行语义标注,并将结果通过 XML 结构化文档进行组织,保留多粒度标注结果的结构层次。在语义标注过程中需要进一步根据需要进行概念描述、语义关系揭示等,其中还包括标注词中的非叙词的描述等。

4.1 学术期刊文本资源标注粒度划分

通过学术期刊的结构特征分析,本文将学术期刊文本资源的标注粒度层次做了划分,如图 3 所示:

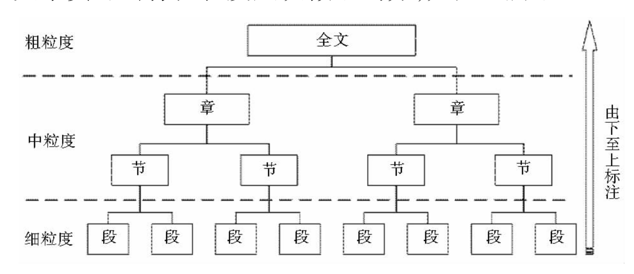


图 3 学术期刊文本资源标注粒度层次划分

依据学术期刊文本资源的主题结构特征,本文在标注过程中的对学术期刊文本资源的标注粒度划分如下:①粗粒度:学术期刊全文内容;②中粒度:学术期刊章、节单位;③细粒度:学术期刊自然段落。

在学术期刊文本资源中,较粗粒度的内容包含了其下进一步划分出的较细粒度的内容,二者在标注过程中往往存在一定的等级关系,因此本文选择由下至上的标注方向,即从学术期刊文本资源的自然段落先进行标注,再标注章节等中粒度内容,最后以全文为单位进行粗粒度标注。

4.2 学术期刊文本资源的多粒度分词

本文将使用由中国科学院开发的汉语分词系统 NLPIR 进行分词。NLPIR 分词工具用到的字典主要有词典库 coreDict、词与词间的关联库 BigramDict、人名库 nr、翻译人名库 tr、地名库 ns。学术期刊的多粒度分词过程是在学术期刊文本资源的粒度划分基础上由下至上,从底层段落开始对各个粒度单位逐一依次分词。分词过程可表示为如图 4 所示:

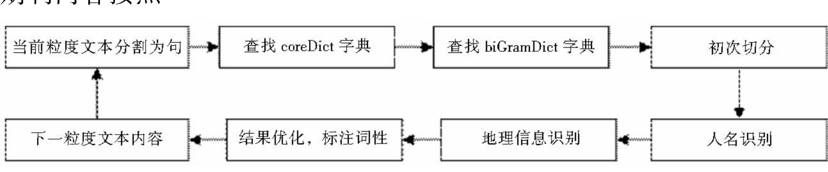


图 4 基于 NLPIR 分词工具的多粒度分词过程

结合图 3 中的粒度划分结构,本文的多粒度分词过程大致如下:首先取细粒度单位文本“段”使用 NLPIR 工具进行分词,按照图 4 中所示 NLPIR 工具的

分词流程完成分词,分词完成后将各个段落的分词结果分别存储并做好标识;然后再取中粒度单位文本“节”,再次按照以上分词流程完成分词,将各节的分词结果分别存储并做好标识。中粒度文本“章”的分析处理流程与文本“节”相同;最后取粗粒度文本“全文”按照以上分词流程完成分词,进行存储并做好标识。多粒度的分词是为了区别分词的文本对象基数不同时对分词算法与分词结果造成的细微差距,如分词中因粒度文本的不同产生的“新词发现”结果的差异,因此往往某一章节下各个“段”的分词结果的总和并不能完全等同与该章节文本的一次性分词结果。通过多粒度的分词可以全面把握当前粒度单位文本的全貌,从而得到最能合适表达当前粒度文本内容的分词结果,为下一步标注词的选取做好准备工作。

4.3 学术期刊文本资源多粒度标注词的选取

在进行标注词的选取计算之前还应对多粒度的分词结果进行预处理工作,其中最重要的即是去除停用词。去除停用词能排除无意义的高频率词对标注结果的影响。去除停用词后就得到具有标注意义的候选词。最后对候选词进行重要性得分计算即可得出标注词。本方法在候选词的重要性得分计算中使用如图 5 所示的以下指标:

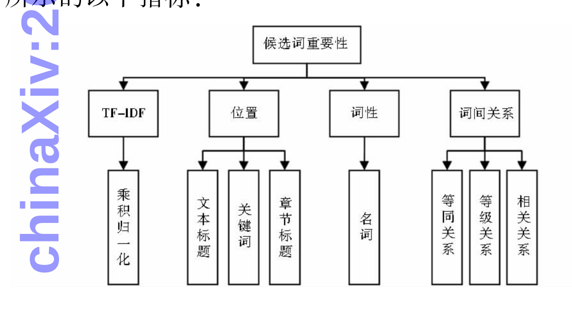


图 5 学术期刊文本资源多粒度标注候选词重要性计算指标

(1)TF-IDF 值。不同的粒度层次中,同一候选词的 TF-IDF 的值都应根据粒度单位重新计算。TF 值指该候选词在当前标注粒度单位内容中出现的频率,其计算公式为:

$$TF_i = \frac{当前粒度文本中的出现频次}{当前文本的分词总数} \quad 式(1)$$

同理,不同粒度层次的 IDF 值也不相同。本次在多粒度语义标注的标注词重要性指标中 IDF 的计算根据粒度单位不同也有不同的定义。

细粒度中,以段为单位,某一段中的某一特定词语的 IDF 值表示为:

$$IDF_i = \log \frac{该学术期刊的总段数}{该学术期刊中出现该词语的段数} \quad 式(2)$$

中粒度中,以章为单位,某一章中的某一特定词语的 IDF 值表示为:

$$IDF_j = \log \frac{该学术期刊的总章数}{该学术期刊中出现该词语的章数} \quad 式(3)$$

粗粒度中,以章为单位,某学术期刊文档中的某一特定词语的 IDF 值表示为:

$$IDF_k = \log \frac{检索系统中的学术期刊文档总数}{出现该词语的学术期刊文档数} \quad 式(4)$$

分别计算各个标注粒度单位中的各候选词的 TF、IDF 值后,二者相乘即可得词语对应的绝对 TF-IDF 值,然后使用“最大最小值”的归一化处理方法将所有的 TF-IDF 值映射到区间[0,1],使数值易于比较。

(2)位置。重要性指标中位置指标采用直接赋值方法,依据位置指标的下级指标设置,若某一候选词出现于“文本标题”“关键词”“章节标题”3 个重要位置,则为其三级指标赋值为 1,3 个三级指标互不干扰,若某候选词同时出现在其中两个及以上位置时可分别赋值,再代入计算公式乘以权重。

(3)词性。在标注过程中的词性筛选可以快速过滤众多不具有标注意义的词语。本方法在对词语的重要性评估指标中经分析设置了“名词词性”的重要性加成,该三级指标依然采用直接赋值方法,即名词性候选词该指标赋值 1,非名词性候选词该指标则为 0,再代入计算公式乘以权重。因此在该计算方法中,当该候选词有多个词性时,则以它在当前粒度单位文本中使用的词性为主,若仅在当前粒度单位文本中就出现多种词性,则有名词词性即可赋值为 1,若无名词词性,由于并不影响其重要性得分计算,则可按照字顺选择。

(4)词间关系。候选词重要性评估指标中的词间关系指较粗粒度单位中的候选词与上一较细粒度的标注词之间的关系,可分为三大类:“等同关系”“等级关系”“相关关系”。本方法使用《汉语主题词表》作为标注使用的情报检索语言,其中的词间关系与评估指标“词间关系”下 3 个评估指标的对应如下:“等同关系”对应《汉语主题词表》中的“用(Y)”和“代(D)”;“等级关系”对应“属(S)”“分(F)”和“族(Z)”;“相关关系”对应“参(C)”。

该重要性评估指标依然采用直接赋值方法，当较粗粒度单位的标注候选词与上一层次较细粒度单位的标注词之间存在有"等级关系""等同关系""相关关系"其中一种时即给该三级指标赋值 1。其中，就当前候选词与上一层次较细粒度单位的某一特定标注词而言，3 种关系互不交叉，即在两个特定的词间只能选择关系表达最为准确的一个词间关系指标赋值，从而避免重复赋值；另一方面，就当前候选词与上一层较细粒度单位的所有标注词而言，则可与不同标注词具有多种类型的关系，但一种类型的关系仅赋值一次，不多次累加赋值。其中需要特别说明的是，在最细粒度的标注中由于不存在更细粒度的标注结果，所有标注候选词的此项指标得分均为 0。

用变量 TI 表示当前候选词在当前标注粒度文本中的 TF-IDF 值，在不同标注粒度文本中同一个候选词具有不同的 TF-IDF 值，应当在每个粒度标注中重新计算；变量 DT、KW、CT 分别表示当前候选词是否出现在文本标题、关键词、章节标题位置，若当前候选词出现在该位置则赋值为 1，否则赋值为 0；变量 N 表示当前候选词是否为名词，若是名词则赋值为 1，否则赋值为 0，当该候选词有多个词性时则以它在当前粒度单位文本中使用的词性为主，若仅在当前粒度单位文本中就出现多种词性，则有名词词性即可赋值为 1；变量 E、G、R 分别表示当前候选词是否与上一层次粒度标注词具有等同关系、等级关系、相关关系，若有则赋值为 1，否则赋值为 0，候选词与某一特定词间仅能选择表达最合适的一种词间关系，不与其他关系重复赋值，而就候选词所在的整个词汇关系网络而言，该候选词则可具备所有类型的词间关系，但这些关系不重复累加赋值。那么候选词 i 在当前标注粒度文本中的重要性得分则可由公式（5）得出：

$$I_i = \frac{TI_i - MIN(TI)}{MAX(TI) - MIN(TI)} + \frac{1}{4} + (DT_i + KW_i + CT_i) * \frac{1}{12} + N_i * \frac{1}{4} + (E_i + G_i + R_i) * \frac{1}{12} \qquad 式(5)$$

通过公式（5）可以计算得出在当前标注粒度文本中各个候选词在"TF-IDF""位置""词性"、"词间关系"4 个方面的综合重要性得分，将得分由高至低降序排序，根据标注需要即可选取适当数量的候选词作为当前粒度文本的标注词。结合学术期刊的多粒度划分结构，由下至上依次完成各个粒度文本的标注词选取，最终构成与学术期刊粒度层次结构对应的多粒度标注结果。

4.4　学术期刊文本资源多粒度标注结果的表示

SKOS 是基于 RDF 的描述语言，它们的基本格式均采用了 XML 格式，因此本文在学术期刊文本资源多粒度标注结果的表示中继续使用 XML 语言进行描述，这样一方面不会与 SKOS 资源描述框架产生冲突，可直接嵌套使用，另一方面 XML 的可扩展性通过自定义标签可以方便地定义标注结果的多粒度层次，保留结构信息。

XML 可扩展标记语言允许用户自定义标签标识结构化文档文容，本文使用 3 个标签组 < document > </document >、< chapter > </chapter >、< paragraph > </paragraph > 来分别标识粗粒度、中粒度、细粒度标注层次。多粒度标注结果的文档结构如图 6 所示：

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdf="http://www.w3.org/2004/02/skos/core#">

<document>
<skos:Concpet rdf:about="粗粒度标注词"/>
<chapter>
<skos:Concpet rdf:about="粗粒度标注词"/>
<paragraph>
<skos:Concpet rdf:about="细粒度标注词"/>
</paragraph>
<paragraph>
<skos:Concpet rdf:about="细粒度标注词"/>
</paragraph>
</chapter>
<chapter>
<skos:Concpet rdf:about="中粒度标注词"/>
<paragraph>
<skos:Concpet rdf:about="细粒度标注词"/>
</paragraph>
<paragraph>
<skos:Concpet rdf:about="细粒度标注词"/>
</paragraph>
</chapter>
</document>
</rdf:RDF>
```

图 6　多粒度标注结果的文档结构示意

5　基于 SKOS 的学术期刊文本资源多粒度语义标注实证研究

5.1　实验基本设定

本次实证研究对象选取了《中国图书馆学报》2016 年第 5 期上顾立平学者发表的一篇理论研究型期刊论文《数据治理——图书馆事业的发展机遇》[19]。该篇

期刊论文具有两个显著特点：

（1）单一作者。单一作者的论文保证了学术期刊前后观点的一致性与问题论述的系统性，学术期刊内部的逻辑结构连贯一体，将更好地反映一种独立思维对某一学术问题的完整思考过程，各个粒度间的相关关系更加突出。

（2）结构严谨。该篇期刊论文以"数据治理是图书馆事业的发展机遇"为论点，全文围绕该主题分别从"数据获取治理""数据共享治理""数据重用治理""数据加值治理"4 个子方面展开论述，形成了典型的"总 – 分 – 总"的结构。严谨的内部结构使得粒度划分更加清晰，各粒度间主题更加明显，同一并列粒度单位间差异显著且不同层次粒度单位间关系密切。

因此，该期刊论文既对学术期刊有较好的代表性，也对本方法有较好的适应性，本文将以该文本为标注对象进行基于 SKOS 的学术期刊多粒度语义标注的实证研究，以验证该方法的可行性与标注效果。

5.2 实验对象粒度划分

分析该学术期刊论文的组织结构，并在文本内部逻辑结构的基础上进行粒度划分，依照 4.1 中所构建的标注框架，将该学术期刊论文的粒度划分表示为树状图，如图 7 所示：

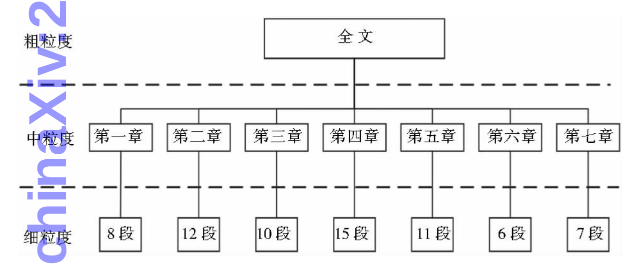


图 7 待标注学术期刊论文文本资源的粒度层次

粒度划分完成后，为了方便接下来的标注工作，还需要对每个粒度单位进行唯一编号。在该等级结构中本方法采用了组合式编号方法，即给每个粒度单位一个"ABC"的编码，其中 A 为全文文档编号，B 为 A 文档中的章节编号，C 为 B 章节中的段落编号，层层递进构成每个粒度单位的唯一编号，A、B、C 的具体表示可根据实际需求进行设定。如在本次实证研究中通过对待标注学术期刊的结构分析，A、B、C、均采用了十进制二位数，那么例如第一章第二段落的文中位置表示即为"010102"。

其中特别的，"标题"位置以"00"表示，如第二章标题位置即为"010200"；全文标题、关键词等不属于任何章节的位置（即在某一划分层级为空）以"00"补缺，并在下一层级的粒度单位继续依次编号，那么全文标题位置即为"010000"，关键词位置即为"010001"。

层级组合式的编号一方面给每一个粒度单位都赋予了唯一的编号，另一方面也十分利于接下来的标注工作中对任一粒度单位内容的读取和标注，如对中粒度的标注时只需区别位置编号中"B"位置即可。

5.3 实验对象多粒度分词

对待标注的学术期刊论文粒度划分完成后，就要分别对各个粒度单位进行分词。使用 NLPIR 分词工具对待标注文本的各个粒度单位内容进行分词，该工具提供了多种分词方法与可供使用的词性标注集，经使用试验文本分词后的结果初步比较，本次实证选择使用其最大匹配的分词方法与 ICTPOS 一级词性标注集。

完成对学术期刊论文的分词后，还需去除分词结果中的停用词，得到标注词的候选词，为每个词语赋予位置信息后导出为 EXCEL 格式，进入下一步的标注词选取计算。其中，停用词表使用由哈尔滨工业大学发布的停用词表扩展版。去除停用词后，部分标注候选词如表 3 所示：

表 3 预处理后部分标注候选词（粗粒度）

词	词性	位置
创新	v	010100
创新驱动	n	010100
引言	n	010100
变革	v	010101
标准	n	010101
财富	n	010101
成就	n	010101
承担	v	010101
持续	v	010101
存储	v	010101
大众	n	010101
反映	v	010101
方式	n	010101
方向	n	010101

5.4 实验对象多粒度标注词选取计算

依照标注候选词重要性评价指标，使用 4.3 节中各指标的得分的计算方法分别计算各个粒度单位文本中的各个标注候选词的各项指标得分，最终得到每个标注候选词的重要性总得分。需要特别说明的是，在本次实证研究中由于难以获取检索系统中的所有文档，因此粗粒度中的 TF-IDF 值暂由 TF 值代替进行组粒度中候选词重要性的计算。

依次由细粒度至粗粒度层次计算各标注候选词的重要性总得分，由高至低排序后根据需要选择标注词的数量。考虑标注词的数量与文本长度相匹配，一般来说，细粒度文本单位以得分最高者为标注词，中粒度文本单位以重要性得分前三位者为标注词，粗粒度文本单位以重要性得分前五位者为标注词。

以中粒度标注单位为例，部分候选词重要性计算结果如表 4 所示：

表 4　部分候选词重要性得分计算结果

词	词性	词性得分	位置	篇标题	节标题	关键词	TF	IDF	TF-IDF	TF-IDF得分	等同关系	等级关系	相关关系	总得分
经济	n	1	01	0	0	0	0.025 540 275	0.243 038 049	0.006 207 259	0.747 722 627	0	0	0	0.436 930 657
创新驱动	n	1	01	0	1	0	0.003 929 273	0.845 098 04	0.003 320 621	0.4	0	0	0	0.433 333 333
数据治理	n	1	01	1	0	1	0.001 964 637	0.146 128 036	0.000 287 088	0.034 582 505	1	0	0	0.508 645 626
数据共享	n	1	01	0	1	1	0.001 964 637	0	0	0	1	0	0	0.5
人类	n	1	01	0	0	0	0.005 893 91	0.845 098 04	0.004 980 931	0.6	0	0	0	0.4
机遇	n	1	01	1	0	0	0.003 929 273	0.544 068 044	0.002 137 792	0.257 517 125	0	0	0	0.397 712 615
引言	n	1	01	0	1	0	0.001 964 637	0.845 098 04	0.001 660 31	0.2	0	0	0	0.383 333 333
价值	n	1	01	0	0	0	0.015 717 092	0.243 038 049	0.003 819 851	0.460 137 001	0	0	0	0.365 034 25
数据驱动	n	1	01	0	1	0	0.003 929 273	0.243 038 049	0.000 954 963	0.115 034 25	0	0	0	0.362 091 896
知识	n	1	01	0	1	0	0.013 752 456	0.066 946 79	0.000 920 683	0.110 904 89	0	0	0	0.361 059 556
服务对象	n	1	01	0	1	0	0.001 964 637	0.367 976 785	0.000 722 941	0.087 084 993	0	0	0	0.355 104 582
查询数据	n	1	01	0	0	0	0.003 929 273	0.845 098 04	0.003 320 621	0.4	0	0	0	0.35
大众	n	1	01	0	0	0	0.003 929 273	0.845 098 04	0.003 320 621	0.4	0	0	0	0.35
地方	n	1	01	0	0	0	0.003 929 273	0.845 098 04	0.003 320 621	0.4	0	0	0	0.35
国家竞争力	n	1	01	0	0	0	0.003 929 273	0.845 098 04	0.003 320 621	0.4	0	0	0	0.35
历史	n	1	01	0	0	0	0.003 929 273	0.845 098 04	0.003 320 621	0.4	0	0	0	0.35
起源	n	1	01	0	0	0	0.003 929 273	0.845 098 04	0.003 320 621	0.4	0	0	0	0.35
潜在价值	n	1	01	0	0	0	0.003 929 273	0.845 098 04	0.003 320 621	0.4	0	0	0	0.35
设施	n	1	01	0	0	0	0.003 929 273	0.845 098 04	0.003 320 621	0.4	1	0	0	0.433 333 333
事业	n	1	01	0	0	0	0.003 929 273	0.845 098 04	0.003 320 621	0.4	0	0	0	0.35
新型知识	n	1	01	0	0	0	0.003 929 273	0.845 098 04	0.003 320 621	0.4	1	0	0	0.433 333 333
序列	n	1	01	0	0	0	0.003 929 273	0.845 098 04	0.003 320 621	0.4	1	0	0	0.433 333 333
时代	n	1	01	0	1	0	0.001 964 637	0.243 038 049	0.000 477 481	0.057 517 125	0	0	0	0.347 712 615
科学	n	1	01	0	0	1	0.031 434 185	0	0	0	0	0	0	0.333 333 333
时间	n	1	01	0	0	0	0.005 893 91	0.367 976 785	0.002 168 822	0.261 254 98	0	0	0	0.315 313 745

从表 4 数据中可以得出该学术期刊的中粒度标注中第一章节的标注词为“经济”“数据治理”“数据共享”。

5.5　实验对象多粒度标注结果

完成各个粒度单位标注候选词的重要性得分计算后，分别为每个粒度单位选取适当数量的候选词作为该粒度单位的标注词。本次细粒度文本单位以得分最高者为标注词，中粒度文本单位以重要性得分前三位者为标注词，粗粒度文本单位以重要性得分前五位者为标注词，该学术期刊各粒度单位的标注结果如表 5 所示：

表 5　多粒度标注词选取结果

粒度层次	文内结构	标注词
粗粒度	全文	数据治理、数据共享、数据重用、数据加值、开放数据
中粒度	第一章	经济、数据治理、数据共享
	第二章	企业家、数据治理、数据加值
	第三章	数据获取治理、数据治理、数据重用
	第四章	生态系统、开放数据、地球
	第五章	数据重用、洪水、开放数据
	第六章	数据加值、数据治理、数据重用
	第七章	图书馆事业、数据治理、数据加值
细粒度	第一章段落	数据共享/战略/特征/经济/序列/设施/报告/新型知识
	第二章段落	科学/步骤/企业家/公民/行业/数据治理/数据工程师/Analyst/管家/角度/数据馆员/需求
	第三章段落	数据获取/政策/开放数据/PLos/数据管理/数据获取治理/搜索引擎/病历记录/利润/哈佛大学
	第四章段落	数据共享/地球/物种/内海/生命/EMIF/药物/患者/人类大脑/注释/效用/载体/数据共享/美国/CODATA/RDA/生态系统
	第五章段落	数据重用/洪水/手机/智能/假设/数据管理/数据重用/DDB/数据重用/同行/CC/开放数据
	第六章段落	数据加值/数据加值/收入/文本/数据加值/技能
	第七章段落	开放获取/课程/数据治理/数据治理/开放许可协议/图书馆事业/动力

将以上标注词使用 SKOS 词汇进行概念描述并使用 XML 结构化文档进行组织，可得到最终该学术期刊的标注 XML 文档。部分标注文档示例如图 8 所示：

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdf="http://www.w3.org/2004/02/skos/core#">
<document>
        <skos:Concept rdf:about="http://www.example.com/concepts#数据共享">
        <skos:prefLabel>Data sharing</skos:prefLabel>
        <skos:related rdf:resource="http://www.example.com/concepts#信息共享">
        </skos:Concept>

        <chapter>
        <skos:Concept rdf:about="http://www.example.com/concepts#经济">
        <skos:prefLabel>Economy</skos:prefLabel>
        <skos:related rdf:resource="http://www.example.com/concepts#部门经济">
        <skos:related rdf:resource="http://www.example.com/concepts#各科经济">
        <skos:related rdf:resource="http://www.example.com/concepts#经济理论">
        <skos:related rdf:resource="http://www.example.com/concepts#经济学">
        <skos:related rdf:resource="http://www.example.com/concepts#区域经济">
        </skos:Concept>

        <paragraph>
        <skos:Concept rdf:about=" http://www.example.com/concepts#战略" >
        <skos:prefLabel>Strategy</skos:prefLabel>
        <skos:altfLabel>军事谋略</skos:altfLabel>
        <skos:altfLabel>战略对策</skos:altfLabel>
        <skos:related rdf:resource=" http://www.example.com/concepts#谋略" >
                </skos:Concept>
</paragrapf>
</chapter>

<chapter>
        <skos:Concept rdf:about=" http://www.example.com/concepts#图书馆事业" >
        <skos:prefLabel>Librarianship</skos:prefLabel>
        <skos:related rdf:resource="http://www.example.com/concepts#比较图书馆学">
        <skos:related rdf:resource="http://www.example.com/concepts#图书馆学">
          </skos:Concept>
        </chapter>

  </document>

</rdf:RDF>
```



图 8　基于 SKOS 的学术期刊论文多粒度语义标注部分 XML 文档示例

5.6　实验结果评估

由于目前还没有任何多粒度标注的应用，因此本次实证研究仅能以当前检索系统内使用的该学术期刊论文的标注结果为对照组，由于二者在标注结果方面不能完全对应，在评估比较过程中本文以检索效果的定性分析作为评估内容。

本文在评估过程中借用“查全”“查准”两个概念对多粒度语义标注效果在相同检索式的情况下与目前使用的标注结果进行理论分析对比，找出多粒度语义标注结果在检索过程中可能对检索系统的查全率与查准率产生的影响。除此之外，检索过程中依据标注结果为用户提供的内部特征的检索入口、检索反馈结果的形式等也是影响用户检索结果和利用效果的重要功能，因此也可作为对多粒度标注结果的参考评估指标。

以该篇学术期刊论文在 CNKI 中的检索情况为参照组，对比分析基于 SKOS 的多粒度标注结果的检索性能，以评估该方法的效用。具体的比较结果如表 6 所示：

表 6　基于 SKOS 的学术期刊论文多粒度标注结果检索性能评估

评估指标	CNKI	多粒度标注
内部特征检索入口	篇名、关键词、摘要、全文	全文主题、章节主题、段落主题
检索结果反馈形式	文档	文档、章节、段落
查全	检索式匹配结果	可通过词间语义关系进行拓展
查准	检索式匹配结果	对文档的内容检索更加准确

（1）从检索系统可能提供的文档内部特征的检索入口来看，目前 CNKI 检索系统中一般提供以自然语言直接匹配的“篇名、关键词、摘要、全文”4 个文档内部特征的检索入口，CNKI 提供的“主题”检索入口是“篇名、关键词、摘要”3 个检索入口的集合，因此不算做单独检索入口。多粒度的语义标注提供了 3 个粒度的内容检索入口，分别为全文主题、章节主题和段落主题。二者提供了不同体系的检索入口，二者不同的检索入口之间相互补充构成整体，很难对比哪种更好，但若以二者其中基本相当的篇名入口和粗粒度检索入口为例，当 CNKI 中以篇名检索该篇学术期刊论文时，以检索词“数据治理”“图书馆”“图书馆事业”“发展机遇”等与篇名“数据治理——图书馆事业的发展机遇”匹配可获得反馈检索结果。在多粒度标注的检索中，以相应的粗粒度为例，其标注词为“数据治理”“数据共享”“数据重用”“数据加值”“开放数据”，那么将以上标注词作为检索词均可获得该篇学术期刊论文。二者相较，前者直接以标题核心词作为该学术期刊论文的主题，后者通过由下至上的多粒度标注后得到该学术期刊论文的主题，实现了对核心主题的扩展，对全文的主题描述更加丰富，但也丢失了如“图书馆事业”这一主题的限制，这一结果有利有弊，在实际的检索过程中则可以在中粒度、细粒度的标注结果中对丢失的主题信息进行补充。

（2）从检索结果的反馈形式来看，目前 CNKI 仅能反馈给检索用户文档单位，即粗粒度内容。多粒度的语义标注结果可对文档进行不同层次结构的组织，在用户检索反馈中可同时呈现不同粒度大小的检索结果，可以是完整文档，可以是某一文档中与检索主题相

关的某一章节,甚至是某一文档中与检索相关的某一具体段落,用户可根据个性化的信息需求进行选择或在检索结果中进行过滤,直接保留某一粒度的检索结果。

(3)从标注结果对检索系统的查全率影响来看,目前 CNKI 中的文档标注结果不能直接提高检索的查全效果,查全率的保障主要依靠检索用户在检索式构建中的检索技巧,如同义词、相关词的扩展等。基于 SKOS 的学术期刊多粒度语义标注结果自身已具备了叙词表中存在的词间关系,可通过等同关系、等级关系、相关关系等语义关系的获取扩大检索范围而不必增加检索用户的智力负担。

(4)从标注结果对检索系统的查准率影响来看,尽管难以判断两种标注结果对检索系统查准率的影响,且在一定程度上来说作为查全率的互逆概念,在检索系统查全率能够明显提升的情况下往往查准率会相应下降。但如果仅考虑检索结果中反馈的主题相关内容,由于多粒度语义标注对文档内部的章节、段落信息都经过了严格的标注处理,在对文档内容的检索方面,若与单纯的字面匹配的全文检索相比则显然会更加准确。

当然,基于 SKOS 的学术期刊多粒度语义标注也将必然带来一些缺点。当完全依赖叙词表对学术期刊进行标注时,必然会出现自然语言与受控语言难以匹配的情况,且叙词表中仅呈现基于学科的概念间的简单关系,无法揭示个性化丰富的概念联系,使得较专业构建的领域本体而言,可以利用的语义关系骤减,甚至有时会增加无关的相关关系,出现标注结果的冗余。此外,多粒度的标注结果较单一粒度明显增多,检索系统的信息处理、存储、检索过程中的匹配计算等都必将更加复杂并产生一些新的问题,检索系统的负担必然加重。

6 总结

本研究以语义标注相关理论为理论基础,以 SKOS 相关技术为技术基础,提出了基于 SKOS 的学术期刊文本资源多粒度语义标注方法并进行了实证研究。该方法主要具有两方面的优势:①SKOS 是目前叙词表描述的最佳方案,较 RDF 与 OWL 语言而言,SKOS 对概念与关系的描述更加灵活且标准化,维护操作简单,基于 SKOS 可以实现多层次的检索,基于叙词表可以实现自动的检索扩展;②多粒度的语义标注可满足用户对不同粒度知识单位的需求,丰富地揭示不同文档同一粒度层次和同一文档不同粒度层次之间的语义关系。

通过实证研究,以当前检索系统的标注结果为参照组,分别从"查全""查准""内部特征检索入口""检索结果反馈形式"4 个方面对比分析了标注结果的优势与缺点。

本研究仅是利用 SKOS 化的叙词表对学术期刊文本资源进行多粒度语义标注的初步尝试,仍有一些问题值得进一步深入研究:本方法仅使用了单一的叙词表作为标注工具,接下来还可以尝试使用多个叙词表进行语义标注,其中将会涉及多个叙词表之间的异构问题、不同体系的概念及语义关系的映射问题、标注中选择词表的优先性问题等。但使用多个叙词表进行语义标注显然可以使标注内容更加丰富,一定程度上能够解决目前一些概念和关系难以描述的问题。此外,基于 SKOS 的学术期刊文本多粒度语义标注的未来应用必然不能离开相关工具、系统、平台的开发。本研究仅仅只是尝试,在实证阶段也仅针对少量样本进行了方法的实验,未来还需在方法进一步完善的基础上开发相关的工具系统,以促进该方法的实际应用。

参考文献:

[1] 余溢文,陈爱萍,赵惠祥. 基于语义网的学术期刊发展初探[J]. 中国科技期刊研究, 2013, 24(5):954-956.
[2] 邱均平,牟楠,楼雯,等. 国内外语义标注研究进展分析[J]. 情报理论与实践, 2014, 37(5):12-16.
[3] W3C. Introduction to SKOS[EB/OL]. [2017-08-03]. https:www.w3.org/2004/02/skos/intro.html.
[4] W3C. Datasets[EB/OL]. [2017-08-03]. https:www.w3.org/2001/sw/wiki/SKOS/Datasets.
[5] 范炜. 语义网环境中的叙词表实例研究——利用SKOS构造机器可理解的知识组织体系[J]. 情报科学, 2006(7):1073-1077.
[6] 贾君枝. 简单知识组织系统与汉语主题词表[J]. 中国图书馆学报, 2008, 34(1):75-78.
[7] 张士男,宋文.《科图法》SKOS 描述方案设计[J]. 现代图书情报技术, 2010, 26(6):7-11.
[8] PASTORSANCHEZ J, MARINTEZ J, RODRIGUEZMUNO V. Advantages of thesaurus representation using the simple knowledge organization system (SKOS) compared with proposed alternatives.[J]. Information research an international electronic journal, 2009, 14(4):422-432.
[9] 王茜,陶兰,王弼佐. 语义 Web 中基于 SKOS 的知识组织模型[J]. 计算机工程与设计, 2007, 28(6):1441-1443.
[10] 熊太纯. 基于 SKOS 的网络信息资源主题标引[J]. 图书馆学研究, 2009(7):63-66.
[11] 朱嘉贤,白伟华,李吉桂. Web 资源的多粒度语义标注及其应

用技术研究[J]. 计算机科学, 2011, 38(8):83-87.

[12] 魏墨济，于涛. 基于领域本体的专业文档语义标注方法[J]. 计算机应用, 2011, 31(8):2138-2142.

[13] 冷伏海，白如江，祝清松. 面向科技文献的混合语义信息抽取方法研究[J]. 图书情报工作, 2013, 57(11):112-119.

[14] CUNNINGHAM H, MAYNARD D, BONTCHEVA K, et al. A framework and graphical development environment for robust NLP tools and applications[C]// Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, Pa, Usa. DBLP, 2002:168-175.

[15] DBpedia. DBpedia home[EB/OL]. [2017-08-05]. http:. wiki. dbpedia. org/.

[16] FERNANDEZ N, FISTEUS J A, FUEENTES D,et al. A Wikipedia -based framework for collaborative semantic annotation[J]. International journal on artificial intelligence tools, 2011, 20(5): 847-886.

[17] 汤怡洁，张敏，丁晓芹. 基于关联数据的集成信息平台语义化实现方法研究[J]. 现代情报, 2016, 36(6):66-73.

[18] 徐绪堪，房道伟，蒋勋,等. 知识组织中知识粒度化表示和规范化研究[J]. 图书情报知识, 2014(6):101-106.

[19] 顾立平.数据治理——图书馆事业的发展机遇[J].中国图书馆学报, 2016, 42(5): 40-56.

作者贡献说明：

夏立新：负责拟题及提出思路，对论文撰写提出修改意见；

郑路：负责论文框架设计，文献采集与撰写，论文修改；

张玉晨：负责文献采集与论文撰写；

翟姗姗：参与论文框架设计，论文修改；

孙晶琼：参与论文框架设计，文稿校对。

A SKOS-based Multi-granularity Semantic Annotation Method for Academic Journal Text Resource

Xia Lixin　Zheng Lu　Zhang Yuchen　Zhai Shanshan　Sun Jingqiong

School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [**Purpose/significance**] Semantic annotation for academic journals is facing two major problems which are how to embody all the concepts in oneontology and mostannotation methods are single granularity semantic annotation models. This paper proposes a SKOS-based multi-granularity semantic annotation method for academic journal text resource, which holds great significance to the development of semantic annotation and provides a method to meetusers' academic information demands. [**Method/process**] With the SKOS description of thesaurus,this paper takes academic journal text for example to achieve a multi-granularity semantic annotation method and carry on an empirical study. [**Result/conclusion**] By setting up the experimental group and the control group respectively, the paper evaluates the annotation effect of the SKOS-based multi-granularity semantic annotation method.

Keywords: semantic annotation　multi-granularity　SKOS thesaurus　academic journal